SPICE+: Evaluation of automatic audio captioning systems with pre-trained language models

Félix Gontier, Romain Serizel, Christophe Cerisara

SPICE+: EVALUATION OF AUTOMATIC AUDIO CAPTIONING SYSTEMS WITH PRE-TRAINED LANGUAGE MODELS

Félix Gontier, Romain Serizel, Christophe Cerisara

Université de Lorraine, CNRS, Inria, Loria, F-54000, France.
felix.gontier@inria.fr, {romain.serizel,christophe.cerisara}@loria.fr

ABSTRACT

Audio captioning aims at describing acoustic scenes with natural language. Systems are currently evaluated by image captioning metrics CIDEr and SPICE. However, recent studies have highlighted a poor correlation of these metrics with human assessments. In this paper, we propose SPICE+, a modification of SPICE that improves caption annotation and comparison with pre-trained language models. The metric parses captions to semantic graphs with a deep dependency annotation model and a refined set of linguistic rules, then compares sentence embeddings of candidate and reference semantic elements. We formulate a score for general-purpose captioning evaluation, that can be tailored to more specific applications. Combined with fluency error detection, the metric achieves competitive performance on the FENSE benchmark, with 84.0% accuracy on AudioCaps and 74.1% on Clotho. Further experiments show that the metric behaves similarly to the full sentence embedding similarity, while the decomposition into semantic elements allows better interpretability of scores and can provide additional information on the properties of captioning systems.

Index Terms— Audio captioning, Evaluation, DCASE

1. INTRODUCTION

Automatic audio captioning (AAC) is the process of describing acoustic scenes with fluent sentences [1]. The task has received increasing attention in recent years, in particular through a dedicated DCASE challenge task[1]. However, the adequacy of current metrics for evaluating AAC systems is an prominent subject of discussion within the community.

Current evaluation metrics are borrowed from the image captioning field. The first metrics, BLEU [2], ROUGE [3], and METEOR [4], originate in machine translation and summarization tasks, and measure the n-gram overlap between candidate and reference sentences. Several studies have subsequently revealed poor correlation of these metrics with human assessments of image caption quality [5, 6, 7]. As a result, captioning-specific alternatives CIDEr [8] and SPICE [9] were proposed. CIDEr retains the approach of n-gram comparison, but integrates reference consensus information and Term Frequency Inverse Document Frequency (TF-IDF) to focus on relevant terms in the description. SPICE instead parses reference captions to a graph containing semantic elements, their attributes, and relations to one another, and evaluates the candidate graph via synonym lemma matching. SPIDEr [10], the combination of SPICE and CIDEr, was found to outperform both in correlation to human assessments, and is now the main metric for evaluating AAC systems.

The properties indicative of acceptable captions are not well defined in the current formulation of the AAC task. Examples from Clotho [11] and AudioCaps [12] associate several human reference captions to examples, each describing events relevant to the annotator instead of the entire scene. Likewise, the same acoustic content can be described with different sentence structure, which reference captions do not cover exhaustively. Evaluation through n-gram overlap, suited for language generation with specific wording, thus seems less relevant for audio captioning. Labbé et al. [13] further note that SPIDEr shows high variability to small differences in caption formulation. As a workaround, they propose SPIDEr-max, where the maximum SPIDEr score out of several candidates produced by the system is retained.

In [14], the authors propose a benchmark to evaluate the accuracy of metrics in choosing the best caption in candidate pairs, against human preference. Most n-gram metrics as well as SPICE perform poorly, cementing the need for further research in caption evaluation. The authors propose FENSE, a metric based on Sentence-BERT [15] embeddings similarity and penalized by fluency error detection. Although this method achieves high correlation with human judgements, Martín-Morató et al. [16] argue that sentence-wise semantic similarity does not necessarily reflect the correspondence of described sounds. Instead, they present CB-score, a summarization-oriented metric that extracts AudioSet tags [17] from captions, then derives a relevance-weighted precision score from the consensus in the reference set. However, the approach is limited to sound classes, and discards their qualifiers and relationships in the description.

In this paper, we propose to combine the semantic graph

This work was funded under the ANR project LEAUDS (Grant No. ANR-18-CE23-0020).

[1] dcase.community/

parsing paradigm of SPICE with the strengths of deep language models. The metric uses a pre-trained language model to produce caption annotations, which are then parsed to semantic graphs with a broad set of linguistic rules. Candidate and reference graphs are compared through embedding similarity. The contributions of this paper are: (i) we identify limitations in the current implementation of SPICE, namely that semantic graphs are incomplete, and candidate evaluation through synonym matching is too restrictive, (ii) we replace components of SPICE with pre-trained models to produce comprehensive semantic representations, and (iii) we propose a score based on the embedding similarity between candidate and reference sub-graphs, that is suitable in the current general task setting but can be adapted to specific sub-tasks.[2]

2. PROPOSED METHOD

The proposed metric evaluates a candidate caption against an arbitrarily large set of references, following the same process as SPICE [9]. First, the caption is annotated with part-of-speech tagging, lemmatization, and dependency parsing, which produces a tree-like representation of grammatical relations in the sentence. The annotation is then parsed into semantic graphs with a set of linguistic rules. Graphs are composed of nodes (nouns), attributes (adjective, verbs) and relations. SPICE views graphs as sets of node, node-attribute, and node-relation-node tuples. The candidate graph is evaluated against a single graph comprising all reference captions by comparing each tuple pair independently.

SPICE+ alters the three processing steps, by performing dependency annotation with a pre-trained language model, adding new linguistic rules and post-processing to semantic graph construction, and evaluating candidates on deep embedding comparison[3].

2.1. Semantic graph parsing

SPICE annotates captions on Universal Dependencies [18] using the Stanford probabilistic context-free grammar (PCFG) parser [19]. From empirical observations on Clotho examples, this parser has a high error rate, which propagates to poor rule-based parsing and leads to incomplete or incorrect semantic graphs. Since then, better parsers have been developed in the natural language processing community [20]. Thus, SPICE+ replaces the original PCFG parser with a language model trained on Universal Dependencies parsing.

A set of linguistic rules is then matched to the dependency annotation to obtain a semantic graph. Five rules are added compared to SPICE to handle formulations often found in audio captions, such as open clausal complements (e.g. "*rain **continues to pour***"), passive nominal subjects (e.g. "*a radio **is being played***"), and clausal noun modifiers (e.g. "*people **talking***"). In addition, two post-processing steps are applied on semantic graphs, that led to better results in empirical experiments. First, coordinating conjunctions systematically duplicate attributes and relations. For instance, "***men and women** are **talking and laughing***" is parsed to two nodes *man* and *woman*, each with both attributes *talk* and *laugh*. Lastly, for expressions in the form "A *of* B" (e.g. "*group of people*", "*body of water*") attributes and relations are associated to node B instead of node A.

2.2. Candidate evaluation

The original implementation of SPICE compares candidate and reference graphs as sets of tuples through synonym comparison. The process outputs a binary matrix, from which true positives, false positives, and false negatives are derive to compute an F-score. This approach discards related but non-synonymous words, which results in a prominent spike at 0 in score distributions [16]. At the same time, in audio captioning there may be several valid ways to describe a sound, e.g. *a metal tube clanging, something hitting a pipe*, that are not exhaustively represented in the reference set.

SPICE+ retains a tuple comparison approach, but replaces binary synonym matching with a continuous similarity measure. First, a pre-trained language model extracts sentence embeddings for each tuple in the candidate and reference graphs. For attribute or relation tuples, the element lemmas are concatenated and fed directly to the model. Then, each candidate embedding X_i is compared to all reference embeddings Y_j through cosine similarity, yielding a similarity matrix $S_{i,j}$. Pseudo-precision and pseudo-recall scores are obtained by averaging maximum similarities along i and j:

$$\hat{P}r = \frac{1}{N_i}\sum_{i=1}^{N_i}\max_j S_{i,j}, \quad \hat{R}e = \frac{1}{N_j}\sum_{j=1}^{N_j}\max_i S_{i,j} \qquad (1)$$

The maximum operation is applied because an element of the candidate is not expected to match all elements of the reference set. The final metric is the pseudo F-score, which is referred as SPICE+emb in experiments.

2.3. Hyper-parameters

Several parameters of the metric can be set to match specific application needs. First, tuples can be weighted in eq. (1), either to account for consensus in the references (e.g. with a relevance measure [16]), or with a pre-defined term-importance dictionary depending on the task focus. The balance between precision and recall in the final score can also be changed to penalize false positives or false negatives respectively. Lastly, a decision on matches can be made by applying a threshold on S. This sets a tolerance on lexical specificity: a high threshold only matches synonymous or identical terms, whereas a lower threshold accepts different but related concepts.

[2]Code is available at: github.com/felixgontier/spice-audio.

[3]Examples on the companion page: felixgontier.github.io/spice-audio

Standalone	AudioCaps					Clotho				
Method	HC	HI	HM	MM	Total	HC	HI	HM	MM	Total
CIDEr	56.2 ± 6.2	96.0 ± 2.4	90.4 ± 3.7	61.2 ± 3.1	71.0 ± 2.2	51.4 ± 6.2	91.8 ± 3.4	70.3 ± 5.7	56.0 ± 3.1	63.2 ± 2.3
SPICE	50.2 ± 6.2	83.8 ± 4.6	77.8 ± 5.2	49.1 ± 3.2	59.7 ± 2.4	44.3 ± 6.2	84.4 ± 4.5	65.5 ± 5.9	48.9 ± 3.1	56.3 ± 2.3
SPIDEr	56.7 ± 6.1	96.0 ± 2.4	90.4 ± 3.7	63.4 ± 3.1	72.2 ± 2.1	53.3 ± 6.2	93.4 ± 3.1	70.3 ± 5.7	57.0 ± 3.1	64.2 ± 2.2
S-BERT (FENSE)	$\mathbf{64.0 \pm 6.0}$	$\mathbf{99.2 \pm 1.1}$	$\mathbf{92.5 \pm 3.3}$	$\mathbf{73.6 \pm 2.8}$	$\mathbf{79.6 \pm 1.9}$	60.0 ± 6.1	$\mathbf{95.5 \pm 2.6}$	75.9 ± 5.3	$\mathbf{66.9 \pm 2.9}$	$\mathbf{71.8 \pm 2.1}$
SPICE+	59.1 ± 6.1	85.4 ± 4.4	83.7 ± 4.6	49.0 ± 3.2	62.0 ± 2.3	46.7 ± 6.2	88.1 ± 4.0	70.3 ± 5.7	48.7 ± 3.1	57.8 ± 2.3
SPICE+emb	63.5 ± 6.0	96.4 ± 2.3	91.6 ± 3.4	70.0 ± 3.0	77.0 ± 2.0	$\mathbf{61.0 \pm 6.0}$	94.7 ± 2.8	$\mathbf{76.3 \pm 5.3}$	61.6 ± 3.0	68.9 ± 2.2
With fluency	AudioCaps					Clotho				
Method	HC	HI	HM	MM	Total	HC	HI	HM	MM	Total
CIDEr	56.2 ± 6.2	95.5 ± 2.6	90.4 ± 3.7	75.4 ± 2.8	78.6 ± 2.0	51.9 ± 6.2	91.8 ± 3.4	76.3 ± 5.3	68.8 ± 2.9	71.3 ± 2.1
SPICE	50.7 ± 6.2	83.4 ± 4.6	78.7 ± 5.1	54.8 ± 3.2	62.8 ± 2.3	44.3 ± 6.2	84.4 ± 4.5	71.1 ± 5.6	57.7 ± 3.1	62.1 ± 2.3
SPIDEr	57.1 ± 6.1	95.5 ± 2.6	90.0 ± 3.7	76.3 ± 2.7	79.1 ± 1.9	53.8 ± 6.2	93.4 ± 3.1	76.7 ± 5.2	68.9 ± 2.9	71.9 ± 2.1
S-BERT (FENSE)	$\mathbf{64.5 \pm 5.9}$	$\mathbf{98.4 \pm 1.6}$	$\mathbf{91.6 \pm 3.4}$	$\mathbf{84.6 \pm 2.3}$	$\mathbf{85.3 \pm 1.7}$	60.5 ± 6.1	$\mathbf{94.7 \pm 2.8}$	80.2 ± 4.9	$\mathbf{72.8 \pm 2.8}$	$\mathbf{75.7 \pm 2.0}$
SPICE+	59.1 ± 6.1	85.0 ± 4.4	84.5 ± 4.5	57.7 ± 3.2	66.8 ± 2.3	46.7 ± 6.2	88.1 ± 4.0	75.4 ± 5.3	56.8 ± 3.1	63.2 ± 2.3
SPICE+emb	63.5 ± 6.0	95.5 ± 2.6	$\mathbf{91.6 \pm 3.4}$	83.4 ± 2.4	84.0 ± 1.8	$\mathbf{61.0 \pm 6.0}$	93.9 ± 3.0	$\mathbf{81.0 \pm 4.9}$	70.0 ± 2.8	74.1 ± 2.1

Table 1. Metric performance on the FENSE benchmark, with 95% confidence intervals and highest accuracy in bold.

3. EXPERIMENTS

3.1. Experimental setup

The FENSE benchmark [14] is used as the main evaluation tool for the proposed metric. The benchmark evaluates the ability of metrics to assess the quality of one candidate caption against another in four cases: human-correct (HC) where both captions are references for the correct example, human-incorrect (HI) where one caption is a reference for a different example, human-machine (HM) where one caption is from a captioning system, and machine-machine (MM) where both captions are produced by a trained system. 250 pairs (921 and 1000 for MM resp.) are formed on AudioCaps and Clotho. Four human annotations of relative quality are collected per pair, and metric performance is given by its accuracy in attributing higher scores to the preferred caption. Note that the benchmark removes all punctuation from Clotho captions, which increases the error rate of dependency annotation.

Experiments consider the general captioning task setting of current datasets. As such, the evaluated metric is the pseudo F-score with default precision-recall balance. Furthermore, no decision threshold or tuple weighting is applied to the similarity matrix S. The dependency and part-of-speech annotations are done with UDify [21]. The Stanford parser [19] is retained for lemma annotations following empirical experimentation. Tuple embeddings are computed with Sentence-BERT [15] (checkpoint `paraphrase-TinyBERT-L6-v2`) to follow FENSE [14]. However, similar results are obtained using MPNet [22].

3.2. Ranking capabilities

The ranking performance of SPICE+ is first evaluated on the FENSE benchmark against current metrics CIDEr, SPICE, and SPIDEr, as well as the FENSE metric based on S-BERT. CB-score [16] is omitted as it is precision-oriented, and is not intended for ranking in the FENSE setting. Table 1 shows the results with or without adding a penalty on fluency using the FENSE error detection model [14].

Author	SPIDEr	Sentence-BERT	SPICE+emb
Xie	**0.319**	0.508	0.530
Zou	0.318	0.494	0.526
Mei	0.309	0.496	0.524
Primus	0.296	0.476	0.529
Kouzelis	0.293	**0.517**	**0.540**
Guan	0.291	0.492	0.523
Kicinski	0.270	0.481	0.512
Pan	0.255	0.474	0.513
Labbe	0.241	0.475	0.508
Baseline	0.224	0.454	0.490

Table 2. DCASE2022 captioning systems evaluation on Clotho-testing. SPIDEr is the challenge ranking metric.

SPICE+ slightly improves on SPICE, indicating that the pre-trained dependency annotator and additional graph parsing rules are beneficial. However, its absolute performance is still poor with less than 50% MM ranking accuracy on both datasets without accounting for fluency. Adding embedding similarity significantly increases performance, confirming the limitations of synonym tuple matching. SPICE+emb also significantly outperforms CIDEr and SPIDEr on MM and total accuracy, the current accepted metrics for audio captioning. Combining SPICE+emb with CIDEr to obtain a SPIDEr equivalent did not improve performance in our experiments.

Overall, SPICE+emb is worse than Sentence-BERT on total accuracy, with a smaller difference when penalizing fluency through error detection. The metric evaluates semantic elements rather than the entire caption, and thus cannot account for fluency. However, this approach provides more information on system performance by analyzing the contributions of elements to the final score, which is not possible with full sentence embedding comparison.

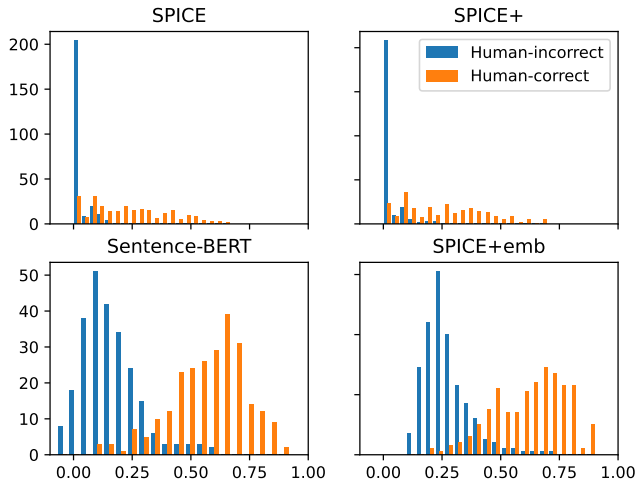

Fig. 1. Distributions of scores for correct and incorrect captions.

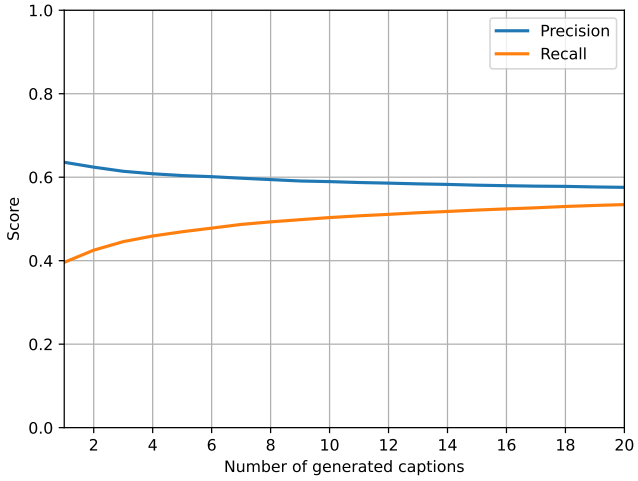

Fig. 2. SPICE+ precision and recall for varying number of candidates generated by the DCASE2022 challenge baseline.

The currently used SPIDEr score, as well as Sentence-BERT and SPICE+emb, are shown in Table 2 for systems of the DCASE2022 challenge on the Clotho-testing dataset[4]. SPICE+emb appears correlated to Sentence-BERT, and significantly alters system ranking compared to SPIDEr.

3.3. Comparison to SPICE behavior

Next, the behavior of SPICE+ on absolute scoring of captions is evaluated. Figure 1 presents the distributions of scores for HI pairs on Clotho, i.e. correct and incorrect captions. For SPICE, most incorrect captions are concentrated at 0, and give little information on the degree of similarity between candidates and references. Interestingly, the distribution is similar for SPICE+ with synonym matching, with only a slight shift of correct caption scores to higher values. In contrast, both embedding-based metrics SPICE+emb and Sentence-BERT provide a bimodal distribution, with some overlap. These metrics reliably assign clear correct and incorrect candidates with high and low scores respectively, which is consistent with the high performance ($> 90\%$) on the HI case in Table 1.

3.4. Captioning system understanding

Systems are conditioned by training datasets to produce sentences with certain properties (e.g. length), which limits their ability to fully describe the acoustic scene in a single caption. When tasked with producing multiple captions, some systems may output variations of the same sentence while others may describe the scene more comprehensively at the cost of lower precision. Because SPICE+ compares semantic elements extracted from captions, it can also enable better understanding the behavior of captioning systems in such conditions.

[4]Full results and technical reports at dcase.community/challenge2022

As a first experiment in this direction, SPICE+emb is computed for the DCASE2022 challenge baseline with varying number of generated captions from 1 to 20, obtained with a beam size of 25. For multiple candidates, semantic graph parsing is symmetrical, i.e. a single graph is constructed from all captions. The pseudo-precision and pseudo-recall tradeoff is shown in Figure 2. Both scores are stable and the recall does not increase to 1, indicating that the system outputs captions with similar content and is unable to match the diversity found in the reference set. SPICE+ is thus able to identify a limitation of the baseline system, and could provide a more comprehensive analysis of captioning systems in this manner.

4. CONCLUSION

We introduced SPICE+, a modification of SPICE including pre-trained language models for dependency annotation and graph comparison. The metric achieves competitive accuracy on a candidate ranking benchmark, but should be used in combination with a fluency measure due to its sensitivity to grammatical errors. Compared to full sentence embedding comparison, SPICE+ is more interpretable as it evaluates individual semantic elements. It is also adaptable through its hyper-parameters to the needs of specific applications that may arise in the community.

Still, some improvements should be investigated in future work. Parts of the captions are currently omitted from the graph, such as adverbs. The similarity between same-graph tuples is not accounted for in the score computation, which gives more importance to nodes with multiple attributes or relations. Lastly, the metric could be specialized to audio by adding a focus on salience and temporal relationships, which are not prevalent in current datasets of short acoustic scenes.

5. REFERENCES

[1] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 374–378.

[2] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[3] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 2004, pp. 605–612.

[4] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *EACL 2014 Workshop on Statistical Machine Translation*, 2014.

[5] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 54, 2013.

[6] D. Elliott and F. Keller, "Comparing automatic evaluation measures for image description," in *52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 452–457.

[7] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby talk: Understanding and generating simple image descriptions," in *CVPR 2011*, 2011, pp. 1601–1608.

[8] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *2015 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *ECCV 2016*, 2016, pp. 382–398.

[10] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 873–881.

[11] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 736–740.

[12] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[13] E. Labbé, T. Pellegrini, and J. Pinquier, "Is my automatic captioning system so bad? spider-max: a metric to consider several caption candidates," in *2022 Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2022.

[14] Z. Zhou, Z. Zhang, X. Xu, Z. Xie, M. Wu, and K. Q. Zhu, "Can audio captions be evaluated with image caption metrics?," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 981–985.

[15] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *EMNLP/IJCNLP*, 2019, pp. 3980–3990.

[16] I. Martín-Morató, M. Harju, and A. Mesaros, "A summarization approach to evaluating audio captioning," in *2022 Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2022.

[17] J. F. Gemmeke, D. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE ICASSP 2017*, 2017.

[18] J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman, "Universal Dependencies v2: An evergrowing multilingual treebank collection," in *12th Language Resources and Evaluation Conference*, 2020, pp. 4034–4043.

[19] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *41st Meeting of the Association for Computational Linguistics*, 2003, pp. 423–430.

[20] D. Zeman, J. Hajič, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, and S. Petrov, "CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies," in *Proc. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2018, pp. 1–21.

[21] D. Kondratyuk and M. Straka, "75 languages, 1 model: Parsing universal dependencies universally," in *EMNLP-IJCNLP*, 2019, pp. 2779–2795.

[22] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *arXiv preprint arXiv:2004.09297*, 2020.